

AWARD NUMBER: W81XWH-14-1-0080

TITLE: Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer.

PRINCIPAL INVESTIGATOR: Christopher B. Umbricht, MD, PhD

CONTRACTING ORGANIZATION: Johns Hopkins University

REPORT DATE: September 2016

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;  
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE September 2016		2. REPORT TYPE Annual		3. DATES COVERED 9/1/2015-8/31/2016	
4. TITLE AND SUBTITLE Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer.				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-14-1-0080	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Christopher B. Umbricht, MD, PhD  E-Mail: cumbrich@jhmi.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Johns Hopkins University  Baltimore, MD 21205				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)  U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT  Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT: This project is designed to complement a multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS that did or did not progress to invasive breast cancer, with an in-depth analysis of expression data on the entire range of informative RNA categories. During the current reporting period, we have completed an Affymetrix HTA 2.0 array-based comprehensive transcriptome assay of samples from 5 collaborating institutions. A gene set enrichment analysis of differentially expressed genes identified statistically significant enrichment of gene sets in both progressive and non-progressive DCIS, including immune-related pathways that may be involved in disease progression. All samples have also undergone a comprehensive DNA methylome analysis using the Illumina 450K CpG arrays, that was successfully used to develop DNA copy number variation (CNV) data on the same cohort. The CNV analysis suggests that Ch8q may harbor a protective genetic marker for DCIS progression, and that chr13 and chr18 deletions were more prevalent in non-progressive DCIS. We have also continued our collaboration with Dr. C. Perou at UNC to maximize the possibility of a successful RNA Sequencing effort, and have obtained encouraging results using the new Illumina TruSeq RNA Access Library Preparation kit followed by RNA sequencing performed using the Illumina HiSeq 2500.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT  Unclassified	18. NUMBER OF PAGES  21	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT  Unclassified	b. ABSTRACT  Unclassified	c. THIS PAGE  Unclassified			19b. TELEPHONE NUMBER (include area code)

## Table of Contents

	<u>Page</u>
<b>1. Introduction.....</b>	<b>2</b>
<b>2. Keywords.....</b>	<b>2</b>
<b>3. Accomplishments.....</b>	<b>2</b>
<b>4. Impact.....</b>	<b>16</b>
<b>5. Changes/Problems.....</b>	<b>16</b>
<b>6. Products.....</b>	<b>18</b>
<b>7. Participants &amp; Other Collaborating Organizations.....</b>	<b>18</b>
<b>8. Special Reporting Requirements.....</b>	<b>18</b>
<b>9. Appendices...(References).....</b>	<b>18</b>

## **1. Introduction.**

Our overall goal remains to develop predictive markers that will be useful in identifying the minority of cases of preinvasive breast cancer (DCIS), that do in fact progress to invasive disease (IBC), and complements our multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS that either progressed to invasive breast cancer IBC (cases) or had no further breast cancer events (controls).

The SPECIFIC AIMS remain essentially unchanged, see section 5 for updated aims:

Specific Aim 1: Perform a descriptive analysis of the DCIS transcriptomes of a multicenter cohort of patients with either progression to invasive breast cancer, or with over 10 years of disease free survival. The objective is to obtain a comprehensive catalog of transcriptome alterations in DCIS, covering differential transcription levels, alternate splicing variation, and non-coding RNA expression (both miRNA and lncRNA), using a state-of-the-art platform.

Specific Aim 2: Perform bioinformatic analyses identifying signatures that are specific for high-risk DCIS, and integrate the transcriptome data with complementary datasets from the same cohort. The objective is to select small sets of features that together discriminate classes, while avoiding over-fitting and benefiting from cross-platform validation.

Specific Aim 3: Develop a panel of multiplex assays that can be used in minimal routine clinical material to predict long-term outcome in DCIS, and optimize performance on in-house DCIS samples. Candidate marker sets will be characterized biochemically and marker-specific assays applicable to high throughput analysis of clinical samples will be developed. Markers that perform well will be combined into multiplex quantitative PCR and Nanostring assays that can be tested for optimal prognostic performance on in house tissue samples.

Specific Aim 4: Validate the results in independent, population-based test cohorts of DCIS patients with progressive disease vs. DCIS patients without recurrent disease, using the newly developed assays. Our objective is to prospectively test our DCIS assay on an independent test set in order to obtain a realistic assessment of its potential positive and negative predictive power.

## **2. Keywords**

Preinvasive breast cancer (DCIS); Invasive breast cancer (IBC); Transcriptome; Prognostic markers; splice variant analysis; non-coding RNA; formalin-fixed paraffin-embedded (FFPE) tissue; Receiver Operator Characteristic (ROC), Area under the Curve (AUC); Estrogen Receptor (ER).

## **3. Accomplishments**

In previous reporting periods, we reported the completion of the accrual, initial characterization and processing of samples from 5 collaborating institutions. We also reported on the successful DNA methylome assessment using Illumina's 450K microarray, and an initial assessment of DNA copy number variation (CNV) based

on a computational method we developed called Epicopy. We have since refined this method to adapt to the specific challenges posed by FFPE-derived DNA. As previously reported, we changed our transcriptome analysis strategy from RNA-sequencing, which yielded insufficiently robust data, to the HTA2.0 microarray from Affymetrix, after obtaining good results in a titration pilot study using a subset of our DCIS samples. Additionally, we have continued our collaboration with UNC to develop a robust next-gen RNA sequencing protocol for DCIS FFPE samples, and have encouraging results from a pilot using the Illumina RNA access analysis pipeline.

## **Transcriptomic analysis of formalin-fixed paraffin-embedded (FFPE) ductal carcinoma in situ (DCIS) using the Affymetrix WT Pico amplification kit and HTA2 gene array.**

### **Methods**

#### **Patient identification and sample collection**

**Table 1: Sample distribution**

	<b>JHH</b>	<b>UAB</b>	<b>UHawaii</b>	<b>UIowa</b>	<b>USC</b>
<b>Discovery Phase</b>					
Case	6	7	1	64	20
Control	6	8	1	54	29
Normal	0	0	0	8	0
<b>Total</b>	<b>12</b>	<b>15</b>	<b>2</b>	<b>126</b>	<b>49</b>

181 DCIS and 8 normal tissue samples passed QC and were used for the analysis.

#### **DNA/RNA co-extraction**

After evaluation of DCIS area by resident pathologist, DCIS epithelial cells were enriched via macrodissection with a clean scalpel. DNA and RNA were extracted using Qiagen (Hilden, Germany) Allprep RNA/DNA FFPE kit with a modified deparaffinization protocol, where samples were deparaffinized in xylene for 3x 10 minute washes instead of manufacturer recommended 10 minute wash.

DNA and RNA yield were quantified using Qubit fluorometer (Qiagen), with the broad range RNA and DNA reagents.

#### **DNA and RNA quality assessment**

DNA quality for methylation profiling was performed using the Illumina (San Diego CA) FFPE DNA QC kit and samples with a delta CT  $\leq 6$  compared to the provided control. RNA quality was assessed using Experion (Biorad, Hercules CA) on a randomly sampled subset of samples with varying yield and age of FFPE block. The RNA assessment showed no significant difference in RIN score or distribution of the RNA fragments across these samples.

#### **Affymetrix HTA2 microarray**

FFPE-derived RNA was processed per manufacturer recommended protocols using the WT Pico kit for global amplification of the RNA and hybridization on the HTA2 microarray. Based on our results from the titration experiment, 10ng total RNA were used as input.

### **Illumina Human Methylation 450k microarray (Illumina 450K array)**

FFPE-derived DNA were restored using the Illumina FFPE DNA restoration kit per manufacturer's recommendation. Restored DNA samples were then hybridized and scanned according to manufacturer provided protocol.

### **Data analysis**

Analyses were performed using the R statistical software [1] with base, Bioconductor [2] and custom functions and packages where necessary.

### **HTA2 data processing**

Per manufacturer recommendation for FFPE-derived RNA, data was processed using the Affymetrix Expression Console using the SST transformation, GCCN correction, and RMA normalization. Batch effects across processing plate were adjusted using COMBAT. Manufacturer recommended QC was performed and the positive vs negative AUC measure of 0.7 was used as a threshold to filter against samples of poor performance and principal component analysis (PCA) was used to identify outliers. A single sample was removed from further analysis, with low positive vs negative AUC and behaving as outlier on PCA analysis.

The Affymetrix HTA-2 Probeset Annotation (Release 36) was used to map probe sets to known genomic features.

### **Illumina 450K data processing**

Raw Idat files of the Illumina 450K array were provided by the SKCCC microarray core and were read using the minfi package. Sample-wise call rate was calculated using a detection p-value cutoff of 1e-05 and density plots were used to evaluate the distribution of beta-values. Samples with <80% call rate or have an aberrant beta-value distribution were excluded from downstream analyses.

Pre-processing was performed using functional normalization. Probe-wise call rate was calculated using a detection p-value cut-off of 1e-05 for all probes, and probes with call rates of < 99% (failed in 2 or more samples) were dropped from the study. Probes within 3 base pairs of a known SNP with 5% minor allele frequency (MAF) were removed from the study.

### **TCGA data**

Processed RNA-seq and Illumina 450K methylation data [3] were obtained from the Firehose GDAC hosted by the Broad Institute, with the data downloaded in August 2015.

### **Differential expression analysis**

Differential expression analysis was performed using linear models for microarray analysis (limma) by constructing a model comparing progressive versus non-progressive DCIS cases.

### **Gene Set Enrichment Analysis**

A rank-based GSEA-like [4] approach was used to perform gene set analysis. Briefly, moderated t-statistics from the DCIS progressive vs. non-progressive limma analysis restricted on RefSeq genes were used to rank the genes. These scores were used to calculate enrichment against the hallmark geneset curated by the Molecular Signatures Database (MSigDB) [5, 6] to identify biologically relevant gene set differences between progressive and non-progressive DCIS.

### **Estrogen receptor (ER)-classification of DCIS samples**

A k top-scoring pairs (KTSP) approach implemented by the switchbox [7] package was used to build an ER classifier for both methylome and transcriptome datasets.

Briefly, an ER classifier was built using invasive breast cancer data obtained from TCGA with unambiguous ER-status using a 10-fold cross validation scheme for parameterization. Two parameters were optimized using cross-validation approaches: 1) the number of features (genes or probes) in the search space (termed feature number,  $F$ ) and 2) k pairs to use in the classifier ( $k$ ).

#### *Feature number or search space optimization*

Feature number was optimized using a 10-fold cross-validation approach where the ER-positive and ER-negative samples were split proportionally into 10 sets, where 9 sets were used as training sets and the remaining set was used as a validation set. The feature number was optimized by altering the search space to obtain a KTSP score for each of the validation samples and assessing prediction accuracy using an ROC analysis to maximize AUC. The number of pairs,  $k$ , was allowed to vary between 3 (minimum requirement) and the rounded up square root of  $F$ .

#### *k optimization*

Following feature number optimization, the optimal number of k TSPs were identified using a similar schema, where  $k \in \{3 \dots F\}$  was used to maximize the AUC of an ROC analysis in the validation dataset.

#### *Voting scheme*

Since previous measurements of prediction potential of the KTSP classifier was performed using ROC analyses, no thresholds were required for making a prediction call. In the application of this classifier in an unknown dataset, a threshold for classification is necessary. The classifier implements a majority vote in its decision process.

#### *Classifier validation & classification*

The ER classifier is then evaluated for predictive accuracy by using it to classify a subset of DCIS with known ER-status. An empirical threshold for AUC was set at 0.8

for the ability to predict ER-status in these samples to constitute success, before using the same classifier for the rest of the DCIS samples. Following validation, the ER-status for all the DCIS samples was predicted.

### **Copy number analysis in DCIS**

Epicopy was used to obtain copy number information from Illumina 450K data. To adjust for FFPE-derived DNA, a more stringent threshold for minimum probe number per segment and fold change was implemented to obtain high quality segment calls. GISTIC 2.0 was used to identify and quantify recurrent copy number variation (CNV) across all DCIS samples. The meta-analysis results from Rane et al. (2015) [8] were obtained for use in a comparative Manhattan plot as the known CNVs in DCIS. A comparative analysis between progressive and non-progressive DCIS was performed by taking the difference of the frequencies of CNV observed across both groups.

### **Pilot experiment for total RNA-seq in tissue abundant follicular thyroid cancer**

#### *RNA Extraction and Quality Assessment*

Unstained histological slides were macro-dissected to enrich for tumor cells (>75%) using a consecutive H&E section annotated by the study pathologist as reference. RNA was extracted from the samples and DNase treated using the Maxwell(r) 16 LEV RNA FFPE Purification Kit (Promega, Madison WI) following the manufacturers protocol. The resulting RNA was analyzed for UV absorbance wavelength ratios (Nanodrop; 260/230, 260/280) to determine purity and concentration. The amount of RNA was normalized to the DV200 value obtained from the Agilent RNA Tapestation, representing the fraction of RNA >200bp in that sample. Where necessary, samples were concentrated using sodium acetate/ethanol precipitation to have a DV200-normalized input of 1ug RNA in 10uL.

RNA fragment distribution was analyzed by the Tapestation and found to be highly degraded, as is expected for FFPE samples, eliminating the need for fragmentation before library preparation.

#### *Library preparation and sequencing*

Ribosomal RNA (rRNA) depletion was performed using the Ribozero Gold rRNA Depletion Kit (Illumina, San Diego). TruSeq Stranded Total RNA Library Prep Kit (Illumina, San Diego) was used for library preparation following manufacturer's protocols, and performed using Agilent Bravo A automated workstation (Agilent, Wilmington DE). Final libraries were analyzed by Tapestation to determine average fragment size. A normalized pool of all 8 samples was sequenced on Illumina MiSeq sequencer as a final QC measure. Sequencing was performed using Illumina HiSeq 2500 at 2 samples per lane using an 8-lane flowcell to produce approximately 150 million paired-ended sequencing reads of 48 base pairs per sample.

#### *Data processing and analysis*

RNA-SeQC [9] was used to assess quality metrics in resulting reads. Reads were aligned to the Human Genome Reference Consortium build 38 (GRCh38) using Tophat2 [10] and assembled into transcripts using CLASS2 [11]. The collection of transcripts merged across the samples was used to identify differentially expressed genes between metastatic and non-metastatic primary tumors with Cuffdiff2 [12], and to determine differential alternative splicing events (exon skipping, mutually exclusive exons, alternative exon ends) with rMATS [13]. Results were inspected and visually validated using the Integrative Genomics Viewer (IGV) [14].

## **Pilot experiment for RNA access in FFPE DCIS samples**

### *RNA Extraction and Quality Assessment*

The same procedures as described above were used for this pilot study.

### *Library preparation and sequencing*

FFPE-derived RNA was processed per manufacturer recommended protocols using the Illumina TruSeq RNA Access Library Preparation kit for global amplification of the RNA. Since the kit captures coding regions, no rRNA subtraction or poly(A)capture steps are required. The maximum recommended amount of total RNA was used because of the typically low DV200 values observed in the DCIS RNA samples. Sequencing was performed using Illumina HiSeq 2500 at 2 samples per lane using an 8-lane flowcell to produce approximately 150 million paired-ended sequencing reads of 48 base pairs per sample.

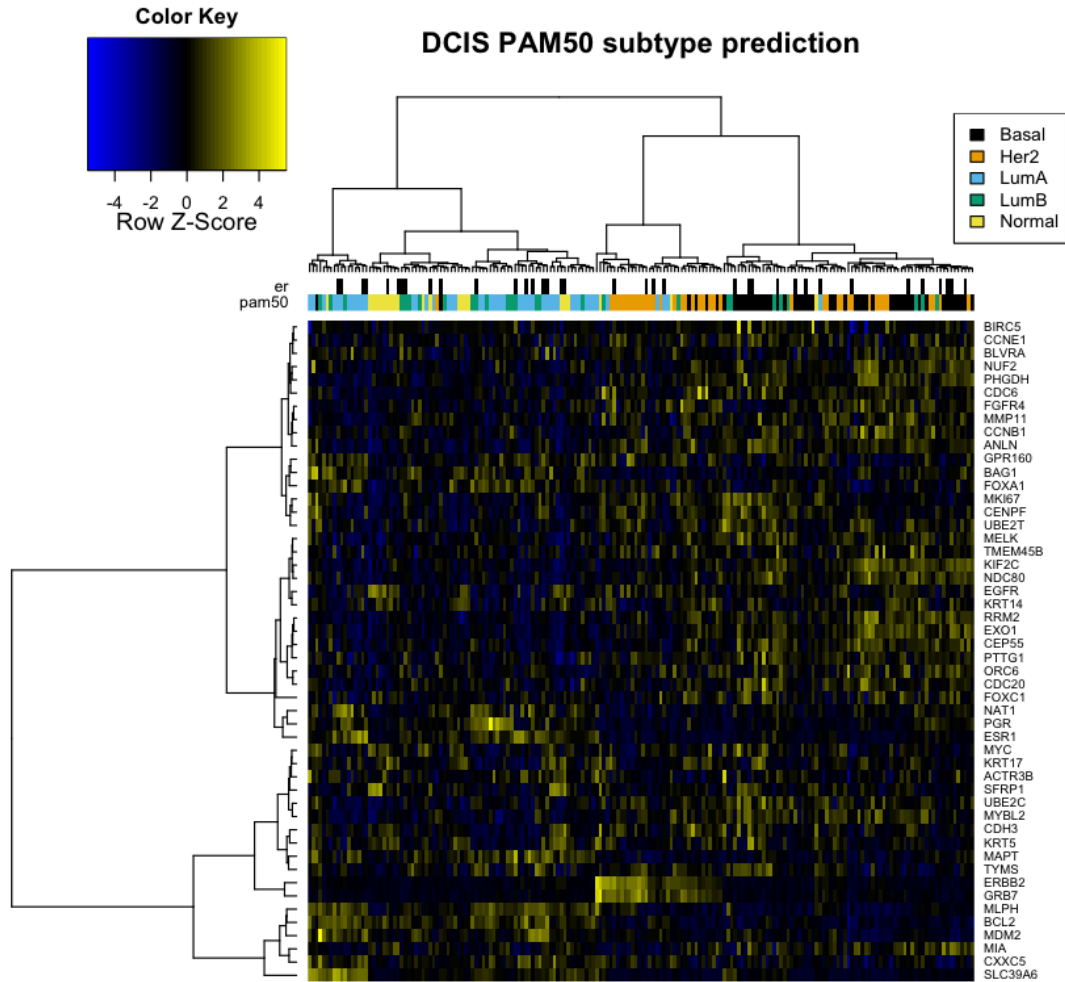
## **Results**

### **Unsupervised analysis**

Unsupervised clustering and principal component analysis (data not shown) revealed that signal intensities were affected by technical variation of unknown origin, which we are actively working to mitigate in collaboration with Affymetrix, the developers of the assay. Despite this limitation, we were able to observe biological signal on subsets of genes.

### **PAM50 classification**

Using the “genefu” [15] package from Bioconductor, the PAM50 intrinsic subtypes, were predicted with a scaled single sample predictor (SSP) method (Fig 1).



**Figure 1:** PAM50 subtypes in DCIS samples.

Structure within the PAM50 genes in these DCIS samples mimics structure in PAM50 molecular subtypes of IDC [3]. For example, we observe co-expression of ESR1 and PGR, as well as ERBB2 and GRB7. Despite the current technical limitations of the data, we were able to observe biologically relevant transcriptomic profiles. Furthermore, this may suggest that DCIS, being a non-obligate precursor, may already have manifest transcriptomic changes that may define the final lesion.

The PAM50 subtypes of the DCIS samples in this study is as follows:

**Table 1: PAM50 subtypes in DCIS**

Subtype	DCIS	
	Progressive	Non-progressive
Basal	21	17
Her2	9	6
LumA	17	13
LumB	8	7

<b>Normal</b>	1	5
<b>Undefined</b>	37	47

The molecular subtypes are equally represented across progressive and non-progressive subtypes.

### Differential expression analysis

The technical effects on the data did not segregate by progression status, and we therefore hypothesize that we should be able to measure biological differences, albeit possibly with a decrease in sensitivity. At this stage, differential expression analysis revealed little difference between progressive and non-progressive disease (Table 2).

**Table 2: Limma results comparing progressive vs. non-progressive DCIS**

<b>HUGO_symbol</b>	<b>logFC</b>	<b>AveExpr</b>	<b>t</b>	<b>P.Value</b>	<b>FDR</b>
ZADH2	-0.06	5.52	-3.76	2.23E-04	1.00
CAPN13	-0.09	4.24	-3.53	5.18E-04	1.00
PGM3	-0.05	4.33	-3.47	6.43E-04	1.00
MAGI2	0.06	4.20	3.38	8.88E-04	1.00
BCL10	-0.05	4.01	-3.37	9.14E-04	1.00
COBL1	-0.05	4.50	-3.30	1.14E-03	1.00
ADSSL1	-0.05	4.43	-3.25	1.37E-03	1.00
MTOR	0.03	4.67	3.15	1.88E-03	1.00
KIF3A	-0.06	4.77	-3.15	1.92E-03	1.00
C11orf53	0.07	3.63	3.13	2.00E-03	1.00
PANX3	0.07	4.08	3.10	2.23E-03	1.00
SNRNP48	-0.06	4.18	-3.09	2.34E-03	1.00
SIX6	0.04	3.52	3.06	2.52E-03	1.00
CNTF	0.06	4.96	3.06	2.53E-03	1.00
NPY2R	0.05	4.23	3.03	2.81E-03	1.00

### Gene set enrichment analysis

We hypothesize that a gene set analysis will allow us to identify biologically significant pathways through this set of samples because 1) pathways could act through multiple genes, and 2) requiring concerted change in a series of genes is a more reliable metric than a single gene.

A GSEA-based pathway analysis approach revealed statistically significant (FDR < 0.05) enrichment of gene sets in both progressive (Table 3) and non-progressive (Table 4) DCIS samples.

**Table 3: Gene sets enriched in progressive DCIS**

Gene Set	p-value	FDR	N
HALLMARK_ALLOGRAFT_REJECTION	1.00E-05	0.00061	197
HALLMARK_KRAS_SIGNALING_DN	0.00086	0.02151	183

**Table 4: Gene sets enriched in non-progressive DCIS**

Gene Set	p-value	FDR	N
HALLMARK_ADIPOGENESIS	0	6.00E-05	191
HALLMARK_FATTY_ACID_METABOLISM	1.00E-05	0.00026	154
HALLMARK_CHOLESTEROL_HOMEOSTASIS	2.00E-05	3.00E-04	73
HALLMARK_ESTROGEN_RESPONSE_EARLY	4.00E-05	0.00049	195
HALLMARK_ANDROGEN_RESPONSE	0.00023	0.00225	99
HALLMARK_ESTROGEN_RESPONSE_LATE	0.00057	0.00478	197

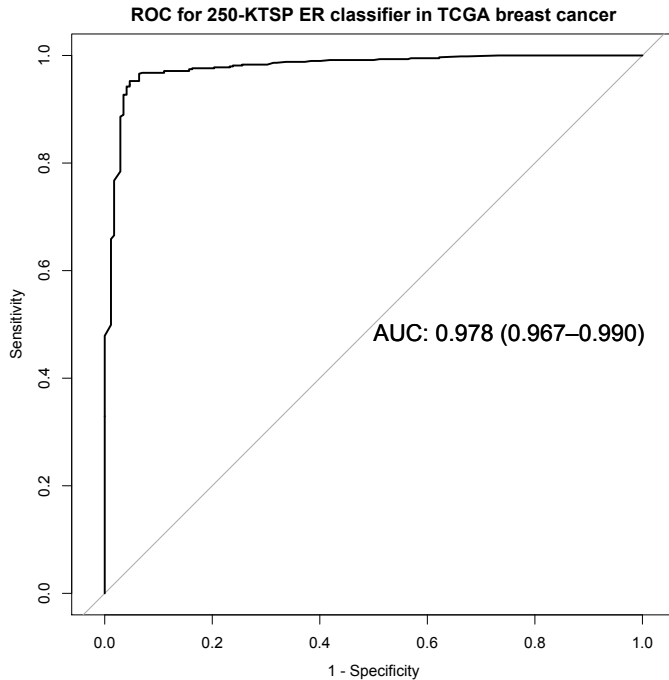
Interestingly, allograft rejection is enriched for in progressive samples, which may suggest for enrichment of immune-related pathways in the progressive samples. In the non-progressive samples, we observed enrichment for estrogen receptor pathways, which may suggest enrichment for ER-positive disease in this group of non-progressive patients.

## ER classification

A KTSP ER classifier was built using TCGA breast cancer (BRCA) RNA-sequencing data. 10-fold cross-validation identified starting feature number as 500 and the maximal number of k's as 250. Using these parameters, a 250-TSP classifier was built, and had an AUC of 0.978 (95% CI 0.967 – 0.990, Fig 2) in being able to predict ER status within the training set. Using majority vote as a threshold, this classifier

achieved an accuracy of 0.957 in the TCGA BRCA dataset.

**Figure 2:** ROC analysis of 250-TSP classifier in the TCGA BRCA training set. The overall accuracy at majority vote is 0.957.



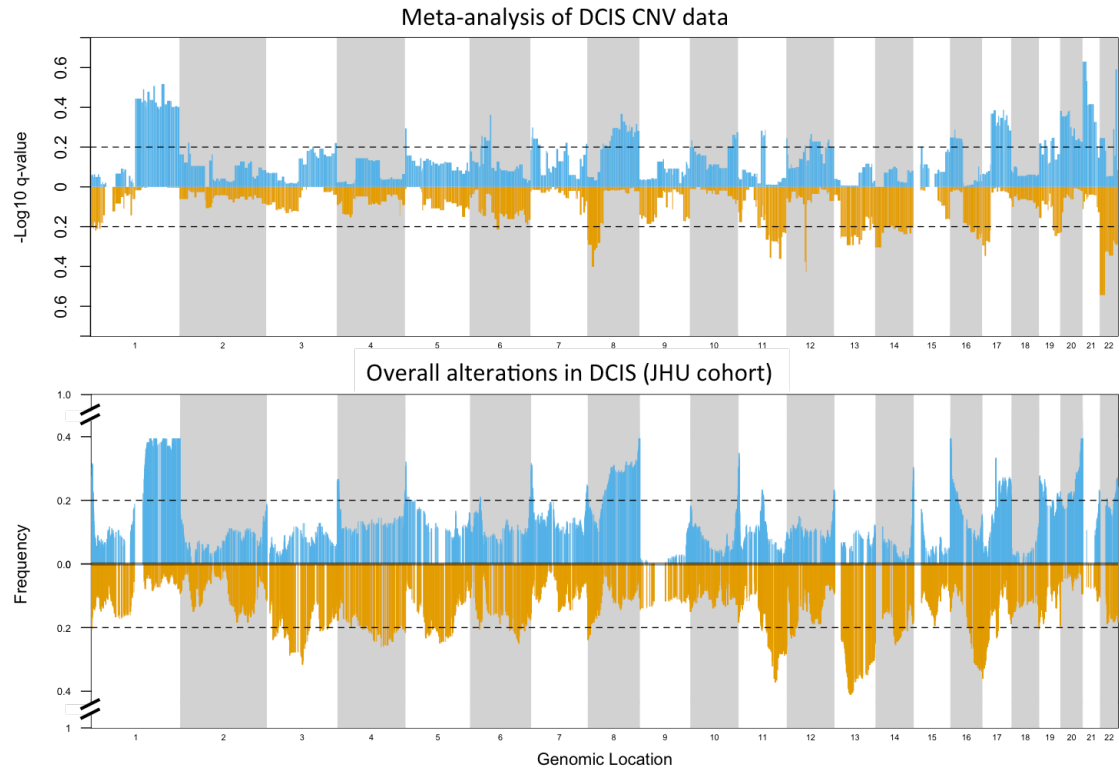
A subset of the DCIS samples had known ER-status, and was used as the validation set for this 250-TSP classifier, with an accuracy of 0.854 (Table 5). Interestingly, when we compared the predicted ER status to the ER-status of the paired IDC of progressive DCIS samples, we observed a concordance of 0.913 (Table 5). This may suggest that progressive DCIS may already develop molecular pathways involved in the progressive disease and would imply that markers of progression in the final disease is detectable in the initial precursor.

**Table 5: Classification metrics of predicting ER-positivity in DCIS samples**

<b>Metric</b>	<b>DCIS: DCIS</b>	<b>DCIS: IDC</b>
Sensitivity	0.971	1
Specificity	0.286	0.5
Accuracy	0.854	0.913

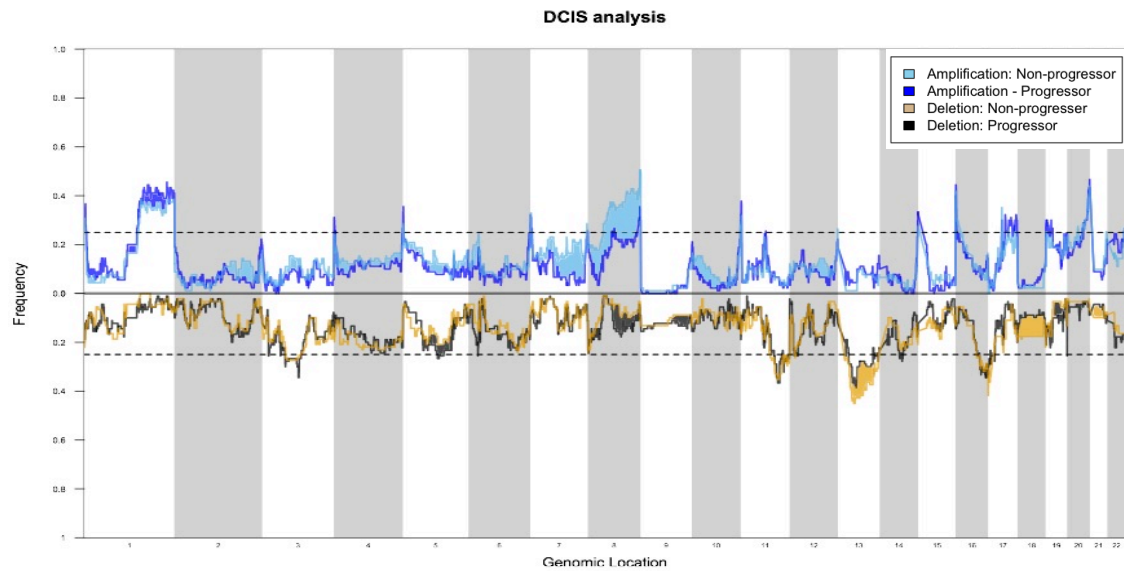
### **Copy number analysis optimized for FFPE materials**

Copy number analysis was also obtained from these samples using Epicopy, a computational pipeline developed by our lab, to augment the information from the transcriptomic data we obtained from HTA2. Since FFPE DNA is of poorer quality compared to fresh frozen (FF) material, we used more stringent parameters for making segment calls. GISTIC 2.0 was used to quantify CNVs in the DCIS dataset, and was compared to known CNVs in DCIS curated by Rane et al. (2015). Many of the recurrent CNVs observed in previous studies were also identified in our DCIS series, suggesting that we were able to detect biologically relevant CNVs using Epicopy with optimized parameters for datasets derived from FFPE (Fig 3).



**Figure 3:** Manhattan plots comparing known recurrent CNV (top) and CNVs identified in the JHU DCIS cohort (bottom), revealing many common CNVs.

We calculated the differences of the frequencies of CNVs across both progressive and non-progressive DCIS.



**Figure 4:** Differences of CNVs across progressive and non-progressive DCIS.

We observed high frequency of amplification of chr8q in non-progressive DCIS (Fig 4). Conversely, there is a higher frequency of deletion of the same region in the progressive DCIS. This behavior suggests that chr8q may harbor a protective genetic marker for DCIS progression. In addition, we also observed chr13 and chr18 deletions being more prevalent in non-progressive DCIS.

### **Pilot study of total RNA-seq in FFPE FTC tissue**

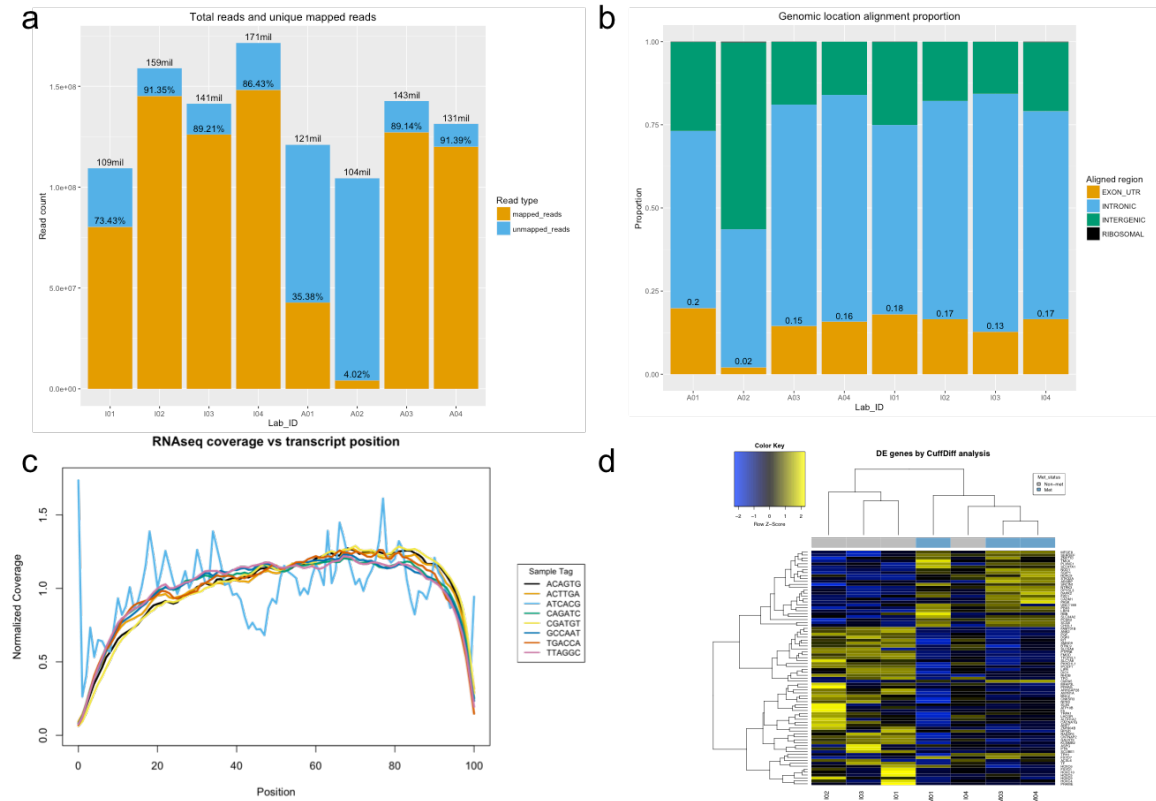
We continue to optimize methods for performing RNA-seq in FFPE material. As a pilot experiment we performed total RNA-seq on a series of follicular thyroid cancer (FTC) FFPE samples. Due to the high cellularity of this tumor type, large amounts of RNA are recoverable for future experiments. Briefly, we compared gene expression analysis of 4 metastatic and 4 non-metastatic primary FTC samples.

#### ***Quality control***

Quality metrics were assessed using RNA-SeQC (Fig 5, a – c). We observed on average 134.9 million reads across 8 samples with an average mapping to the genome of 79.4%, with the exception of a single sample with 4.02% (Fig 5a). In the seven samples with reasonable genomic mapping, 16.3% mapped to exons, which reflects metrics observed in FF RNA-seq samples (Fig 5b). Finally, we observed no transcript position bias in the RNA-seq reads (Fig 5c).

#### ***Differential expression and gene set analysis***

Differential expression analysis revealed 140 differentially expressed transcripts between metastatic and non-metastatic samples. We were able to further identify a single non-metastatic sample, I04, clustered with the metastatic samples, and upon updating of the clinical information, we identified that the patient from whom this sample was derived developed metastatic 10.5 years post initial diagnosis. In other words, by performing total RNA-seq on a small series of samples, we were able to predict metastasis in an FTC sample 10 years before the metastasis presented itself. Furthermore, a GSEA-based gene set analysis revealed enrichment for gene sets involved in metastasis and aggressive behavior (Table 6), including epithelial mesenchymal transition (EMT).



**Figure 5:** Results for RNA-seq pilot experiment in 8 FTC samples. a) Total and unique mapped reads. b) Genomic alignment to different regions in the human genome. c) Transcript position bias analysis. d) Hierarchical clustering of differential expression analysis results using CuffDiff2.

**Table 5: Gene set analysis comparing metastatic and non-metastatic FTC**

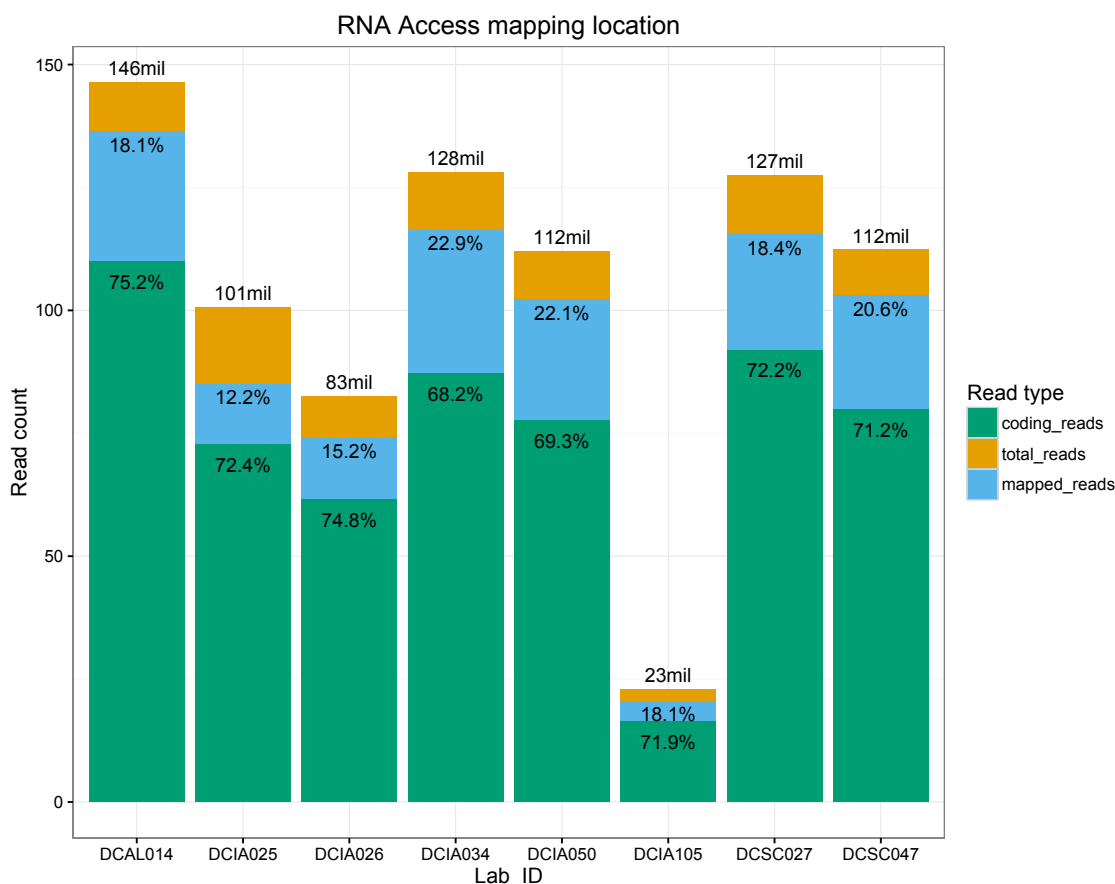
Gene_set	P_value	FDR	N
EPITHELIAL_MESENCHYMAL_TRANSITION	0	0	176
ESTROGEN_RESPONSE_EARLY	0	0	175
NOTCH_SIGNALING	1.00E-05	7.00E-05	26
ADIPOGENESIS	1.00E-05	7.00E-05	164
P53_PATHWAY	2.00E-05	0.00019	172
TNFA_SIGNALING_VIA_NFKB	5.00E-05	0.00038	179
ESTROGEN_RESPONSE_LATE	0.00013	0.00092	180
UV_RESPONSE_DN	0.00015	0.00095	131
OXIDATIVE_PHOSPHORYLATION	2.00E-04	0.00114	163
HYPOXIA	0.00115	0.00577	182

## Pilot study of RNA Access targeted RNA-seq in FFPE DCIS tissue

As a pilot study, we performed RNA Access on a subset of 8 DCIS samples.

### *Reads and mapping rate*

Of the 8 samples, 7 samples had an average total read count of 115.7 million reads and a single sample with 23 million reads (Fig 6). Interestingly, the average percent of reads that mapped to coding regions were 74.8% (range 68.2% to 75.2%), including the sample with poor read counts at 71.9%.



**Figure 6:** Mapping of RNA Access targeted sequencing results in the pilot study.

## Conclusion

Taken together, we conclude that despite technical challenges that still persist in the data, we were able to detect biologically relevant signals from these DCIS samples, which we can augment with high quality methylation and copy number data. Furthermore, our continued assessment and improvement of RNA-seq technologies yielded increasingly promising results, suggesting the possibility of applying one of these technologies to the current set of samples.

#### 4. Impact

N/A

#### 5. Changes/Problems

See discussion of our results in section 3.

##### **Updated Specific Aims:**

**Aim 1: Develop novel methods of assessing quality of samples and performing normalization across FFPE samples of variable quality.**

**Aim 1a: Apply more stringent quality control parameters for enrichment of sample with high quality data.**

We will apply more stringent quality control metrics, such as positive vs. negative AUC, to filter against poor quality samples. Furthermore, unsupervised hierarchical clustering will be performed to further identify samples that cluster with poor quality samples, which will be analyzed separately or filtered away. This will enrich for samples affect by less technical variability, which would allow us to perform exploratory data analysis on a series of samples with good quality data.

**Aim 1b: Optimize thresholds of qRT-PCR-based QC analysis of FFPE samples for identification of samples that will yield reproducible data.**

We also plan to perform qRT-PCR-based QC analysis of a subset of the DCIS samples, selected across a range of HTA2 quality metrics. We will assess correlations between this additional technical measure of quality and some of the HTA2 quality metrics. These results will be published and used to guide future experimental cohorts, including for downstream RNA-seq experiments.

**Aim 1c: Integrate transcriptomic, methylome, and copy number data to identify biomarkers of progression in DCIS samples.**

We expect our collaboration with Affymetrix to yield additional insights into the technical effect on the transcriptomic data. Furthermore, we anticipate the possibility of using custom probe library definitions that will allow for better normalization of the data.

**Aim 2: Perform multi-omic analysis of transcriptome, methylome, and copy number data of DCIS.**

**Aim 2a: Develop novel approaches, including non-parametric methods, to analyzing FFPE data with variable quality.**

We plan to optimize appropriate methods and novel pipelines to analyzing FFPE HTA2 data of variable quality. The use of non-parametric, or parameter free, methods such as KTSP to build phenotypic classifiers and additional rank based algorithms for comparing progressive and non-progressive DCIS that is less affected by data quality. Non-parametric measures of distance, such as Spearman distances, will be used in cluster analyses.

**Aim 2b: Identify subtypes across DCIS samples and learn molecular alterations unique to those subtypes across all three molecular platforms through exploratory data analysis.**

We will use various integrative clustering approaches, such as iCluster, and novel applications of consensus cluster plus or non-negative matrix factorization, to explore the data and identify clusters. Features contributing to each cluster will be used in gene set analysis to determine biologically relevant pathways unique to each cluster. Furthermore, concerted changes of a series of gene across signatures and pathways would be more robust measures, especially in this dataset with technical variability.

Lastly, we will apply additional methods such as multiple concerted disruption to identify the most likely pathological pathway contributing to disease and study interactions across different molecular platforms.

**Aim 2c: Integrate transcriptomic, methylome, and copy number data to identify biomarkers of progression in DCIS samples.**

Using optimized methods from aims 2a and 2b, we aim to identify biomarkers of progression across all three datasets. We will integrate these biomarkers into a panel of cross-platform markers, and perform feature selection using various methods such as Akaike's Information Criterion or penalize lasso regression. Methods that incorporate molecular understanding of the disease such as multiple concerted disruption will also be used to prioritize features.

**Aim 3: Perform RNA Access on a subset of DCIS samples, which allows for both comparative assessment of RNA species across methodologies and technical validation of genes of interest**

**Aim 3a: Perform sample-to-sample assessment of HTA2 and RNA Access data to identify commonalities, as well as differences across platforms.**

RNA Access targeted RNA-seq will be performed on a subset of the DCIS samples stratified for progression and ER-status. This will allow us to compare transcriptomic data across both platforms and assess the reproducibility of specific transcripts in either platform. Furthermore, we will be able to use the data as technical validation of gene signatures of interest. Beyond that, such a comparison will also allow us to assess the limitations of each platform.

#### **Aim 4: Validate genes of interest and biomarkers.**

##### **Aim 4a: Develop bench-based assays and perform technical validation on a phenotypically-stratified subset of DCIS samples.**

Cost-effective bench-based assays will be developed for biomarkers distinguishing progressive and non-progressive disease. Depending on the final number of selected features, different assay technologies will be used. On a limited set of features, QPCR (SYBR green or Taqman probes) or digital PCR-based assays will be developed. On a larger set of features, we will assess the use of technologies such as Nanostring or targeted sequencing if it minimizes cost.

These assays will be used for technical validation of a phenotypically-stratified subset of DCIS samples.

##### **Aim 4b: External validation of biomarkers in DCIS validation cohort.**

Once we observe technical validation of DCIS samples used in the previous study, we will extend that technology for external validation in an additional set of DCIS samples.

## **6. Products**

N/A

## **7. Participants & Other Collaborating Organizations**

Charles M. Perou, Ph.D, The May Goldman Shaw Distinguished  
Professor of Molecular Oncology Departments of Genetics, and  
Pathology & Laboratory Medicine  
Lineberger Comprehensive Cancer Center  
125 Mason Farm Road  
The University of North Carolina at Chapel Hill Chapel Hill, NC  
27599

## **8. Special Reporting Requirements**

N/A

## **9. Appendices**

## References

1. R-Core-Team: **R: A language and environment for statistical computing.**: R Foundation for Statistical Computing; 2016.
2. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
3. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**:61-70.
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550.
5. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection.** *Cell Syst* 2015, **1**:417-425.
6. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 2011, **27**:1739-1740.
7. Marchionni L, Afsari B, Geman D, Leek JT: **A simple and reproducible breast cancer prognostic test.** *BMC Genomics* 2013, **14**:336.
8. Rane SU, Mirza H, Grigoriadis A, Pinder SE: **Selection and evolution in the genomic landscape of copy number alterations in ductal carcinoma in situ (DCIS) and its progression to invasive carcinoma of ductal/no special type: a meta-analysis.** *Breast Cancer Res Treat* 2015, **153**:101-121.
9. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: **RNA-SeQC: RNA-seq metrics for quality control and process optimization.** *Bioinformatics* 2012, **28**:1530-1532.
10. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14**:R36.
11. Song L, Sabunciyany S, Florea L: **CLASS2: accurate and efficient splice variant annotation from RNA-seq reads.** *Nucleic Acids Res* 2016, **44**:e98.
12. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2013, **31**:46-53.
13. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y: **rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.** *Proc Natl Acad Sci U S A* 2014, **111**:E5593-5601.
14. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform* 2013, **14**:178-192.
15. Gendoo DM, Ratanasirigulchai N, Schroder MS, Pare L, Parker JS, Prat A, Haibe-Kains B: **Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer.** *Bioinformatics* 2016, **32**:1097-1099.